Influence of observations on the misclassification probability in quadratic discriminant analysis

Christophe Croux* Kristel Joossens *

K.U. Leuven K.U. Leuven

Abstract

In this paper it is studied how observations in the training sample affect the misclassification probability of a quadratic discriminant rule. An approach based on partial influence functions is followed. It allows to quantify the effect of observations in the training sample on the performance of the associated classification rule. Focus is on the effect of outliers on the misclassification rate, merely than on the estimates of the parameters of the quadratic discriminant rule. The expression for the partial influence function is then used to construct a diagnostic tool for detecting influential observations. Applications on real data sets are provided.

Keywords: Classification, Diagnostics, Misclassification Probability, Outliers, Partial Influence Functions, Quadratic Discriminant Analysis.

*Dept. of Applied Economics, K.U. Leuven, Naamsestraat 69, B-3000 Leuven, Belgium, Email: christophe.croux@econ.kuleuven.ac.be; kristel.joossens@econ.kuleuven.ac.be

1 Introduction

In discriminant analysis one observes two groups of multivariate observations, forming together the *training sample*. For the data in this training sample, it is known to which group they belong. On the basis of the training sample a discriminant function Q will be constructed. Such a rule is used afterwards to classify new observations, for which the group membership is unknown, into one of the two groups. Data are generated by two different distributions, having densities $f_1(x)$ and $f_2(x)$. The higher the value of Q the more likely the new observation has been generated by the first distribution. Taking the log-ratio of the densities yields

$$Q(x) = \log \frac{f_1(x)}{f_2(x)}.$$

For f_1 a normal density with mean μ_1 and covariance matrix Σ_2, and for f_2 another normal density with parameters μ_2 and Σ_2, one gets

$$Q(x) = \frac{1}{2}\left\{(x-\mu_2)^t\Sigma_2^{-1}(x-\mu_2) - (x-\mu_1)^t\Sigma_1^{-1}(x-\mu_1)\right\} + \frac{1}{2}\log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right). \qquad (1.1)$$

Here, $|\Sigma|$ stands for the determinant of a square matrix Σ. The above equation can be written as a quadratic form

$$Q(x) = x^tAx + b^tx + c, \qquad (1.2)$$

where

$$A = \frac{1}{2}(\Sigma_2^{-1} - \Sigma_1^{-1}) \qquad (1.3)$$

$$b = \Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2 \qquad (1.4)$$

$$c = \frac{1}{2}\log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) + \frac{1}{2}(\mu_2^t\Sigma_2^{-1}\mu_2 - \mu_1^t\Sigma_1^{-1}\mu_1). \qquad (1.5)$$

The function $Q(x)$ is called the quadratic discriminant function. Although it has been derived from normal densities it can also be applied as such without making distributional assumptions.

Future observations will now be classified according to the following discriminant rule: if $Q(x) > \tau$, where τ is a selected cut-off value, then assign x to the first group. On the other hand if $Q(x) < \tau$, then assign x to the second group. Now let π_1 be the prior probability that an observation to classify will be generated by the first distribution, and set $\pi_2 = 1-\pi_2$. For normal source distributions it is known that the optimal discriminant rule, in the sense of minimizing the expected probability of misclassification, is given by the above quadratic rule with $\tau = \log(\pi_2/\pi_1)$, e.g. Johnson and Wichern (2002, Chapter 11). In practice, the prior probabilities π_1 and π_2 are often unknown and one uses $\tau = 0$.

The discriminant function (1.1) still depends on the unknown population quantities μ_1, μ_2, Σ_1 and Σ_2, and needs to be estimated from the training sample. So let $x_1, \ldots, x_{n_1}$ be a sample of p-variate observations coming from the first distribution H_1^0 and $x_{n_1+1}, \ldots, x_n$ a second sample drawn from H_2^0. These samples together constitute the training sample. An observation in the training sample will influence the sample estimates of location and covariance, and hence the discriminant rule. In Quadratic Discriminant Analysis (QDA) the primary interest is not in knowing or interpreting the parameter values in (1.2). The aim is to use QDA for classification purposes. Focus in this paper is on how observations belonging to the training sample affect the total probability of misclassification, and this effect will be quantified by the influence function. Influence functions in the multi-sample setting were already considered by several authors, e.g. Fung (1992,1996b). In this paper, the formalism of partial influence functions (Pires and Branco, 2002) as an extension of the traditional influence function concept to the multi-sample setting will be followed.

In the case of equal covariance matrices $\Sigma_1 = \Sigma_2 = \Sigma$ the linear discriminant rule of Fisher results as a special case of (1.1):

$$L(x) = (\mu_1 - \mu_2)^t \Sigma^{-1} (x - \frac{\mu_1 + \mu_2}{2}). \tag{1.6}$$

Influence analysis for Linear Discriminant Analysis (LDA) has been studied by Campbell (1978), Critchley and Vitiello (1991) and Fung (1992, 1995a). The quadratic case seems to be much harder. Some numerical experiments have been conducted to assess the influence of outliers in the training sample on QDA (e.g. Lachenbruch, 1979), while Fung (1996a) proposes several influence measures based on the leave-one-out approach. A more formal approach to influence analysis for quadratic discriminant analysis seems not to exist yet in the literature.

In Section 2 of the paper, a population expression for the total probability of misclassification is presented. The latter is then used as a starting point to compute the partial influence functions for the classification errors in Section 3. The expressions obtained for the partial influence function are not only valid when the classical sample averages $\hat{\mu}_1$, $\hat{\mu}_2$ and sample covariance matrices $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ are used to estimate the unknown population parameters in the discriminant function Q, but also when robust estimators are used. Computations are tedious here and most details have been moved to the Appendix. Besides being of theoretical interest, measuring the influence of an observation in the training sample on the future classification error can be used as a diagnostic tool to detect influential observations. Section 4 presents such a diagnostic tool for diagnosing influential points in a classical discriminant analysis, based on the usual sample averages and covariances. However, to make this diagnostic measure robust, i.e. not suspect to masking effects, robust estimates of the population parameters need to be plugged in the theoretical expressions of the influence functions. Several examples in Section 4 illustrate the use of this diagnostic tool. Finally, some conclusions are made in Section 5.

2 Total Probability of Misclassification

In this Section a population version of the Total Probability of Misclassification (TPM) will be presented. Denote $H^0 = (H_1^0, H_2^0)$, where H_1^0 and H_2^0 are the distributions having generated the training samples. The population version of the quadratic discriminant rule is then, by analogy with (1.2),

$$Q(x; H^0) = x^t A(H^0)x + b(H^0)^t x + c(H^0), \tag{2.1}$$

where the population values of the coefficient of the discriminant rule are

$$\begin{aligned}
A(H^0) &= \frac{1}{2}\left\{C_2(H^0)^{-1} - C_1(H^0)^{-1}\right\} && (2.2)\\
b(H^0) &= C_1(H^0)^{-1}T_1(H^0) - C_2(H^0)^{-1}T_2(H^0) && (2.3)\\
c(H^0) &= \frac{1}{2}\log\left(\frac{|C_2(H^0)|}{|C_1(H^0)|}\right) && (2.4)\\
&\quad + \frac{1}{2}\left\{T_2(H^0)^t C_2(H^0)^{-1}T_2(H^0) - T_1(H^0)^t C_1(H^0)^{-1}T_1(H^0)\right\}.
\end{aligned}$$

In the above formula $T_1(H^0)$ and $T_2(H^0)$ are the values of a location functional T at the distributions H_1^0 and H_2^0. When performing classical discriminant analysis one gets the population averages, i.e. $T_1(H^0) = E_{H_1^0}(X)$ and $T_2(H^0) = E_{H_2^0}(X)$. Similarly, $C_1(H^0)$ and $C_2(H^0)$ are the values of a scatter matrix functional C at the distributions H_1^0 and H_2^0. For classical discriminant analysis, C yields the population covariance matrix, i.e. $C_1(H^0) = \mathrm{Cov}_{H_1^0}(X)$ and $C_2(H^0) = \mathrm{Cov}_{H_2^0}(X)$. In this paper, focus is on *classical* quadratic discriminant analysis, where one uses the conventional population averages and population covariances, resulting in $Q = Q_{Cl}$. However, it is also possible to use *robust* measures of location for T and robust measures of scatter for C, yielding a different discriminant rule denoted by Q_R. For information on robust estimators of location and scatter we refer to Hampel et al. (1986) and Maronna and Yohai (1998).

The distribution generating the future data is supposed to be a normal mixture $H = \pi_1 H_1 + \pi_2 H_2$, with $H_1 = N_p(\mu_1, \Sigma_1)$ and $H_2 = N_p(\mu_2, \Sigma_2)$. The probability of classifying observations from the first group in the second is given by

$$\Pi_{2|1}(H^0, H) = P(Q(X; H^0) < 0 \mid X \sim H_1), \tag{2.5}$$

and the probability of misclassification for observations following H_2 is

$$\Pi_{1|2}(H^0, H) = P(Q(X; H^0) > 0 \mid X \sim H_2).$$

The total probability of misclassification, or the error rate for classifying observations from H using a discriminant rule Q estimated from H^0, is then defined as

$$\text{TPM}(H^0, H) = \pi_1 \Pi_{2|1}(H^0, H) + \pi_2 \Pi_{1|2}(H^0, H). \tag{2.6}$$

If we want to emphasize that we work with the classical discriminant rule Q_{Cl}, we will use the notation TPM_{Cl}. It is important to distinguish between H^0 and H. In the above definitions, no parametric assumptions are made on the distribution generating the training data. The quadratic discriminant rule can be applied to any data set, although it might be expected that the rule performs poor if the data are far from normally distributed. For example, they might contain a few outliers. However, to compute a misclassification rate for future data, a parametric assumption is needed to obtain computable expressions. The normality assumption on H is taken here and the results obtained in this paper all make use of this assumption. The next proposition gives an expression for the TPM.

Proposition 1. *With the notations above, for $H = \pi_1 N_p(\mu_1, \Sigma_1) + \pi_2 N_p(\mu_2, \Sigma_2)$, and for the quadratic discriminant rule $Q(X; H^0)$ defined in* (2.1),

$$\Pi_{2|1}(H^0, H) = P\left(\sum_{j=1}^{p} \lambda_j (W_j - d_{2|1}^t v_j)^2 < k\right) \tag{2.7}$$

where $W_1, \ldots, W_p$ are i.i.d. univariate standard normal. Furthermore, $d_{2|1}$ is a p-variate vector given by

$$d_{2|1} = d_{2|1}(H^0, H) = \Sigma_1^{-1/2} \left(-\frac{1}{2} A(H^0)^{-1} b(H^0) - \mu_1 \right), \tag{2.8}$$

$$k = k(H^0) = \frac{1}{4} b(H^0)^t A(H^0)^{-1} b(H^0) - c(H^0), \tag{2.9}$$

and $\lambda_j = \lambda_j(H^0, H)$ and $v_j = v_j(H^0, H)$ are the eigenvalues and eigenvectors of the matrix

$$\bar{A}_{2|1}(H^0, H) = \Sigma_1^{1/2} A(H^0) \Sigma_1^{1/2}. \tag{2.10}$$

The expression for $\Pi_{1|2}(H^0, H)$ is given by

$$\Pi_{1|2}(H^0, H) = P\left(\sum_{j=1}^{p} \lambda_j (W_j - d_{1|2}^t v_j)^2 > k \right) \tag{2.11}$$

with λ_j and v_j now the eigenvalues and eigenvectors of $\bar{A}_{1|2}(H^0, H)$. Here, $d_{1|2}(H^0, H)$ and $\bar{A}_{1|2}(H^0, H)$ are given by replacing the index 1 by 2 in the definitions of $d_{2|1}(H^0, H)$ and $\bar{A}_{2|1}(H^0, H)$. The total probability of misclassification is then $\mathrm{TPM}(H^0, H) = \pi_1 \Pi_{2|1}(H^0, H) + \pi_2 \Pi_{1|2}(H^0, H)$.

When performing a discriminant analysis, one expects that the data to be classified come from the same distribution as the training data, although the proportions of data coming from the first or second group may be different. In this case, where $H_0 = (H_0^1, H_0^2) = (H_1, H_2)$, we say that we the training data follow the model distribution (and in particular contain no outliers). So at the model, the training data follow a normal distribution as well and $T_1(H^0) = \mu_1$, $T_2(H^0) = \mu_2$, $C_1(H^0) = \Sigma_1$ and $C_2(H^0) = \Sigma_2$. (When we work with Q_R instead of Q_{Cl}, we require consistency of the robust location and covariance measures at the normal distribution.) Hence at the model, the total probability of misclassification is a function of the population parameters of location and covariance. Numerical computation of this TPM requires evaluation of the cumulative distribution function of a linear combination of p chi-squared distributions with one degree of freedom. Note that some of the weights λ_j in this linear combination appearing in (2.7) may be negative, since they are eigenvalues of the symmetric, but in general not positive definite matrix (2.10). Using modern computing power, (2.7) can equally easy be computed with Monte-Carlo integration techniques. Indeed, for a sufficiently high number of vectors $(W_1, \ldots, W_p)$ generated from a multivariate standard normal distribution, we check for every simulated vector whether the inequality in (2.7) holds for the given value of k. The probability in (2.7) is then being approximated as the corresponding empirical frequency.

For diagonal covariance matrices and $H^0 = H$, an expression of the TPM for QDA was presented by Houshmand (1993). Recently, McFarland and Richards (2002) considered the problem of computing exact misclassification probabilities in the normal case for finite samples. The expression for TPM in the setting of Linear Discriminant Analysis is much better known. In the normality case with equal covariances it is simply given $\mathrm{TPM}_{\mathrm{LDA}} = \Phi(\frac{-\Delta}{2})$, with $\Delta = \sqrt{(\mu_1 - \mu_2)^t \Sigma^{-1} (\mu_1 - \mu_2)}$ the Mahalanobis distance between the populations and Φ the c.d.f. of a standard normal. To study the effect of outliers on the total probability of misclassification, partial influence functions will be computed in the next section.

3 Partial Influence Functions

Influence functions have already been used for estimators that depend on more than one sample (e.g. Campbell, 1978; Fung, 1992, 1996b). We compute the influence of observations in the training sample on the TPM by using the formalism of partial influence functions (Pires and Branco, 2002). Partial influence functions (PIF) extend the traditional concept of influence function to the multi-sample setting. The first PIF gives the influence on the classification error of an observation x being allocated to the first group of training data. The second PIF measures the influence on the TPM for training data being allocated to the second group. Formally,

$$\mathrm{PIF}_1(x; \mathrm{TPM}, H^0, H) = \lim_{\varepsilon \downarrow 0} \frac{\mathrm{TPM}\left(((1-\varepsilon)H_1^0 + \varepsilon \Delta_x, H_2^0), H\right) - \mathrm{TPM}(H^0, H)}{\varepsilon}, \tag{3.1}$$

$$\mathrm{PIF}_2(x; \mathrm{TPM}, H^0, H) = \lim_{\varepsilon \downarrow 0} \frac{\mathrm{TPM}\left((H_1^0, (1-\varepsilon)H_2^0 + \varepsilon \Delta_x), H\right) - \mathrm{TPM}(H^0, H)}{\varepsilon}, \tag{3.2}$$

where Δ_x is a Dirac measure putting all its mass at x. One sees that for the first PIF contamination is only induced for H_1^0, the distribution generating the first group of training data, while the second distribution H_2^0 remains unaltered. Only contamination in the training sample is considered, the distribution H of the data to classify is not subject to contamination. When actually computing influence functions, we work at the model distribution $H^0 = (H_1, H_2)$. Indeed, when no contamination is present, one supposes that the data generating processes for the training data and for future data are the same. This model condition is natural and implicitly made in the classification literature. At the model, the notation $\mathrm{PIF}_s(x;\mathrm{TPM},H) := \mathrm{PIF}_s(x;\mathrm{TPM},(H_1,H_2),H)$, for $s = 1,2$, can be used. For classical quadratic discriminant analysis the partial influence functions are written as $\mathrm{PIF}_s(x;\mathrm{TPM}_{Cl},H^0,H)$, for $s = 1,2$. When using robust plug-in estimates in the definition of Q, the notation $\mathrm{PIF}_s(x;\mathrm{TPM}_R,H^0,H)$ is used.

For linear discriminant analysis, the above influence functions have already been computed (e.g. Croux and Dehon, 2001). The result, when using standard population averages and covariances, is strikingly simple

$$\mathrm{PIF}_s(x;\mathrm{TPM}_{Cl}^{\mathrm{LDA}},H^0,H) = (\pi_1 - \pi_2)\frac{\phi(\Delta/2)}{2\Delta}(L(x) - L(\mu_s)) \tag{3.3}$$

for $s = 1,2$. Here ϕ is the density of a standard normal distribution and Δ as before the Mahalanobis distance between the 2 source populations. As Critchley and Vitiello (1991) noticed, the influence is determined by the factor $L(x) - L(\mu_s)$, which they consider as a residual. For QDA it seems very difficult to come up with an easily interpretable expression.

The next proposition shows how the partial influence functions of the TPM using the quadratic discriminant rule Q can be obtained.

Proposition 2. *Let H^0 be the distribution of the training data and $H = \pi_1 N_p(\mu_1,\Sigma_1) + \pi_2 N_p(\mu_2,\Sigma_2)$ the distribution of the data to classify. Suppose that*

(i) All eigenvalues of the matrix $\Sigma_1\Sigma_2^{-1}$ are distinct and different from one.

(ii) The partial influence function of the location functionals T_1 and T_2, and the scatter functionals C_1 and C_2 exist at H^0.

(iii) The model holds, i.e. $H^0 = (H_1,H_2)$.

The partial influence functions of the total probability of misclassification of a quadratic discriminant rule Q based on the location measures $T_1(H^0)$ and $T_2(H^0)$ and the scatter measures $C_1(H^0)$ and $C_2(H^0)$ is then given by

$$\mathrm{PIF}_s(x;\mathrm{TPM},H^0,H) = \pi_1\mathrm{PIF}_s(x;\Pi_{2|1},H^0,H) + \pi_2\mathrm{PIF}_s(x;\Pi_{1|2},H^0,H), \tag{3.4}$$

for $s = 1, 2$. *Here*

$$\begin{aligned}
\mathrm{PIF}_s(x; \Pi_{2|1}, H^0, H) &= \sum_{j=1}^{p} \frac{\partial \Pi_{2|1}(H^0, H)}{\partial \lambda_j} \cdot \mathrm{PIF}_s(x; \lambda_j, H^0, H) \\
&\quad + \sum_{j=1}^{p} \frac{\partial \Pi_{2|1}(H^0, H)}{\partial d_j^*} \cdot \mathrm{PIF}_s(x; d_j^*, H^0, H) \qquad (3.5) \\
&\quad + \frac{\partial \Pi_{2|1}(H^0, H)}{\partial k} \cdot \mathrm{PIF}_s(x; k, H^0, H),
\end{aligned}$$

where the notations of Proposition 1 are used and $d_j^*(H^0, H) = v_j(H^0, H)^t d_{2|1}(H^0, H)$. *Furthermore*

$$\mathrm{PIF}_s(x; \lambda_j, H^0, H) = v_j^t \Sigma_1^{1/2} \mathrm{PIF}_s(x; A, H^0) \Sigma_1^{1/2} v_j, \qquad (3.6)$$

$$\mathrm{PIF}_s(x; d_j^*, H^0, H) = \mathrm{PIF}_s(x; v_j, H^0, H)^t d_{2|1}(H^0, H) + v_j^t \mathrm{PIF}_s(x; d_{2|1}, H^0, H), \qquad (3.7)$$

$$\mathrm{PIF}_s(x; k, H^0) = -\frac{1}{4} b^t A^{-1} \mathrm{PIF}_s(x; A, H^0) A^{-1} b + \frac{1}{2} b^t A^{-1} \mathrm{PIF}_s(x; b, H^0) - \mathrm{PIF}_s(x; c, H^0), \qquad (3.8)$$

while

$$\mathrm{PIF}_s(x; d_{2|1}, H^0, H) = -\frac{1}{2} \Sigma_1^{1/2} \left(A^{-1} \mathrm{PIF}_s(x; b, H^0) - A^{-1} \mathrm{PIF}_s(x; A, H^0) A^{-1} b \right). \qquad (3.9)$$

$$\mathrm{PIF}_s(x; v_j, H^0, H) = \sum_{k=1, k \neq j}^{p} \frac{v_k^t \Sigma_1^{1/2} \mathrm{PIF}_s(x; A, H^0) \Sigma_1^{1/2} v_j}{\lambda_j - \lambda_k} v_k, \qquad (3.10)$$

for $j = 1, \ldots, p$. *The shorthand notations* $A = A(H^0)$, $b = b(H^0)$, $\lambda_j = \lambda_j(H^0, H)$ *and* $v_j = v_j(H^0, H)$ *for* $j = 1, ..., p$, *are used. Furthermore,*

$$\begin{aligned}
\mathrm{PIF}_s(x; A, H^0) &= (-1)^{s+1} \frac{1}{2} \left\{ \Sigma_s^{-1} \mathrm{PIF}_s(x; C_s, H^0) \Sigma_s^{-1} \right\}, \qquad (3.11) \\
\mathrm{PIF}_s(x; b, H^0) &= (-1)^{s+1} \left\{ \Sigma_s^{-1} \mathrm{PIF}_s(x; T_s, H^0) - \Sigma_s^{-1} \mathrm{PIF}_s(x; C_s, H^0) \Sigma_s^{-1} \mu_s \right\}, \qquad (3.12) \\
\mathrm{PIF}_s(x; c, H^0) &= (-1)^{s+1} \frac{1}{2} \left\{ \mu_s^t \Sigma_s^{-1} \mathrm{PIF}_s(x; C_s, H^0) \Sigma_s^{-1} \mu_s \right. \qquad (3.13) \\
&\quad \left. -2 \mu_s^t \Sigma_s^{-1} \mathrm{PIF}_s(x; T_s, H^0) - \mathrm{trace}\left(\Sigma_s^{-1} \mathrm{PIF}_s(x; C_s, H^0) \right) \right\}.
\end{aligned}$$

for $s = 1, 2$. *The partial derivatives* $\frac{\partial \Pi_{2|1}(H^0,H)}{\partial \lambda_j}$, $\frac{\partial \Pi_{2|1}(H^0,H)}{\partial d_j^*}$ *and* $\frac{\partial \Pi_{2|1}(H^0,H)}{\partial k}$, *for* $j = 1, ..., p$, *do not depend on the argument* x, *neither on location and covariance functionals. Expressions for them are given in Lemma's 1, 2 and 3 in the Appendix. In order to compute* $\mathrm{PIF}_s(x; \Pi_{1|2}, H^0, H)$, *it suffices to replace* Σ_1 *by* Σ_2 *in the expressions* (3.6) *up to* (3.10) *and to interchange* $d_{2|1}$ *with* $d_{1|2}$. *The* λ_j *and* v_j *are then the eigenvectors and eigenvalues of the matrix* $\Sigma_2^{1/2} A(H^0) \Sigma_2^{1/2}$ *instead of of the matrix* $\Sigma_1^{1/2} A(H^0) \Sigma_1^{1/2}$.

Computing the partial influence functions appearing in Proposition 2 is tedious, but straightforward. Building bricks are the expressions for the partial influence functions of the estimators of location and scatter. For the classical estimators it is immediate to check that

$$\mathrm{PIF}_s(x; C_s, H^0) = (x - \mu_s)(x - \mu_s)^t - \Sigma_s \text{ and } \mathrm{PIF}_s(x; T_s, H^0) = x - \mu_s, \tag{3.14}$$

for $s = 1, 2$ while $\mathrm{PIF}_s(x; C_{s'}, H^0) = \mathrm{PIF}_s(x; T_{s'}, H^0) = 0$ for $s' \neq s$. From (3.14) all other auxiliary partial influence function can be computed, resulting in $\mathrm{PIF}_1(x; \mathrm{TPM}_{Cl}, H^0, H)$ and $\mathrm{PIF}_2(x; \mathrm{TPM}_{Cl}, H^0, H)$.

Computation of the partial derivatives of $\Pi_{2|1}(H^0, H_1)$, appearing in (3.5), requires some care. These partial derivatives only depend on the population parameters, they do not depend on x, neither on the estimators used. Lemmas 1, 2, and 3 formulated in the Appendix express them in terms of integrals, which can be computed by numerical integration. Note that numerical integration is much more stable than numerical differentiation. Although the formulas for computing the PIF are cumbersome, there are no major computational difficulties. A matlab program computing the partial influence functions is available from *www.econ.kuleuven.ac.be/christophe.croux*.

When deriving the expression for the PIF, the assumption " (i): All eigenvalues of the matrix $\Sigma_1\Sigma_2^{-1}$ are distinct and different from one" was needed. If the matrix $\Sigma_1\Sigma_2^{-1}$, or equivalently $\Sigma_2\Sigma_1^{-1}$, has eigenvalues close to 1, or close to each other, then it can be seen from (3.10) and Lemmas 1 and 2 in the Appendix that the influence function will tend to explode. If one is close to a setting where condition (i) is not valid, then the discriminant rule is very sensitive to single observations in the training data. One case where (i) is not valid is the equal covariance matrix case, where all eigenvalues of $\Sigma_1\Sigma_2^{-1}$ are equal to one. Hence, for reasons of local robustness, it is advised to use LDA whenever one is close to the equal covariance matrix case. Performing a test for equal covariance matrices before carrying out a QDA, as is common in applied research, can prevent construction of an unstable quadratic discriminant rule. However, there are other situations where condition (i) is not met, for example when Σ_1 and Σ_2 are both proportional to the identity matrix. The latter corresponds with a setting of two spherically symmetric data clouds. Here, alternative methods like regularized Gaussian discriminant analysis (Bensmail and Celeux, 1996) are preferable to keep the local sensitivity under control.

The eigenvalues of $\Sigma_1\Sigma_2^{-1}$ determine the nature of the quadratic form (1.2). For example, in the bivariate setting the eigenvalues determine whether the classification regions associated with the two groups are an ellipse and its complement or an hyperbola and its complement. When an eigenvalue passes from below to above one, the nature of the classification regions changes. Finally, note that interchanging two eigenvalues close to each other leads to a change in orientation of the quadratic form, which explains why the equal eigenvalue case is unstable as well (similar as in principal components analysis, see Critchley 1985).

Some pictures of partial influence functions in the univariate and bivariate case are represented. Figure 1 gives the first PIF for $H_1 = N(0,1)$ and $H_2 = N(1,\sigma^2)$, for $\sigma^2 = 0.6, 0.8, 1.2$ and 1.6, and equal prior probabilities for discriminant analysis based on Q_{Cl}. It is immediate to see that the influence functions have a quadratic shape and are unbounded. When the value of σ^2 approaches 1, the values for the PIF increase. For $\sigma^2 = 1.2$ the shape of the PIF is reversed: outliers for the first training data set tend to decrease the estimated error rate.

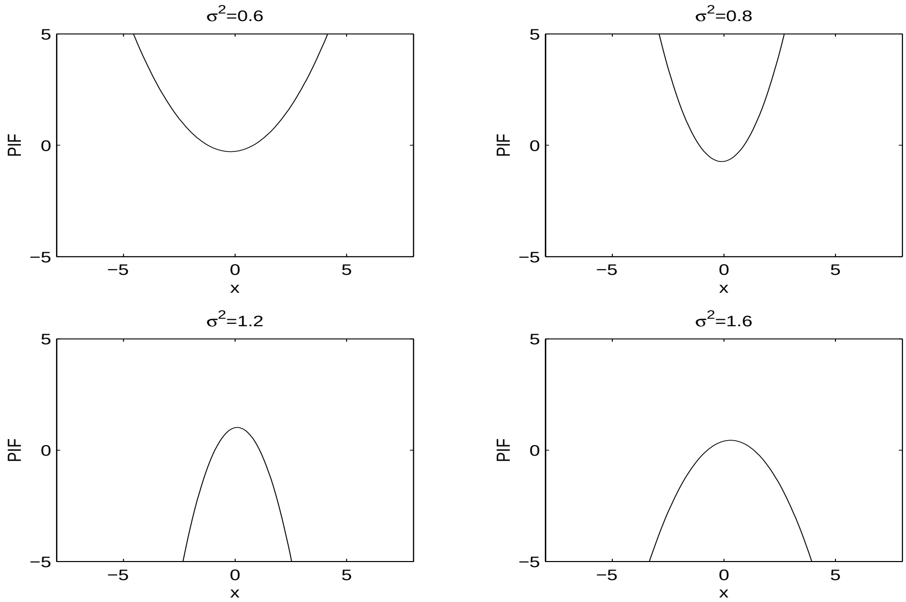

Figure 1: First partial influence function $\mathrm{PIF}_1(x; \mathrm{TPM}_{Cl}, H)$ *for* $H = 0.5N(0,1) + 0.5N(0,\sigma^2)$ *and for several values of* σ^2.

Of course, in practice one is interested in the higher dimensional case. The shape and sign of the PIF depend heavily on the parameter values and are difficult to predict, in contrast with the linear case. In Figure 2 the first partial influence function is shown for a bivariate distribution where $H_1 = N(0, I_2)$ and $H_2 = N((1,1)^t, \text{diag}(0.3, 0.8))$. Notice again the quadratic shape of the influence surface, being quite flat in the central region here, but unbounded in the tails of the distribution.

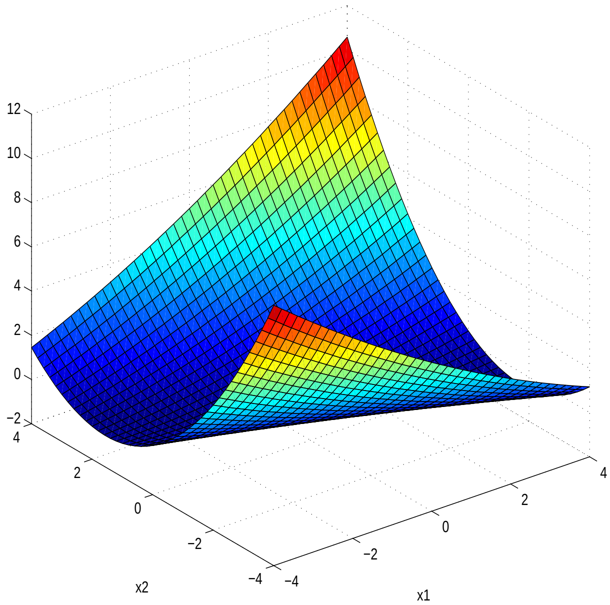

Figure 2: First partial influence function $\text{PIF}_1(x; \text{TPM}_{Cl}, H)$ *for* $H = 0.5N(0, I_2) + 0.5N((1,1)^t, \text{diag}(0.3, 0.8))$.

The expressions in Proposition 2 are not only valid for TPM_{Cl}, but they also apply when robust estimators are used for the parameters μ_1, μ_2, Σ_1 and Σ_2 in the discriminant rule Q. For example, Randles et al. (1978) proposed to use M-estimators. Since M-estimators loose robustness when the dimension p increases, we will use the highly robust Minimum Covariance Determinant (MCD) estimator (Rousseeuw and Van Driessen, 1999). The MCD-estimator is obtained by selecting the subsample of size h (we selected $h = 0.75n$) for which the determinant of the covariance matrix computed from that subsample is minimal, and computing afterwards the mean and the sample covariance matrix solely from this "optimal" subsample. The robustness of the MCD-estimator in the context of QDA has recently been shown by means of simulation studies (Joossens and Croux 2004; Hubert and Van Driessen, 2004). Now, using the results of Proposition 2, we are able to prove local robustness by means of partial influence functions. It is indeed immediate to see that $\text{PIF}_s(x; \text{TPM}, H^0, H)$ is bounded as soon as $\text{PIF}_s(x; \mu_s, H^0)$ and $\text{PIF}_s(x; \Sigma_s, H^0)$ are bounded. Influence functions for the MCD-estimator where computed by Butler, Davies and Jhun (1993) and Croux and Haesbroeck (1999) and were shown to be bounded at elliptical models.

Figure 3 shows the PIF for the same distributions as for Figure 1, but now using the robust MCD estimator to estimate the discriminant rule. The same scaling of the axes as in Figure 1 is used, and it is immediately observed how much lower the values for the PIF become. In the central part of the data, the PIF behaves like the PIF of the classical estimation procedure, but in the tails we observe a bounded influence. Hence far outliers receive a bounded, but non-zero, influence. Notice that for σ^2 close to 1, where condition C is not valid, the influence function also gets blown up, but to a much lesser degree. For σ^2 equal to one, the PIF will not exist either.

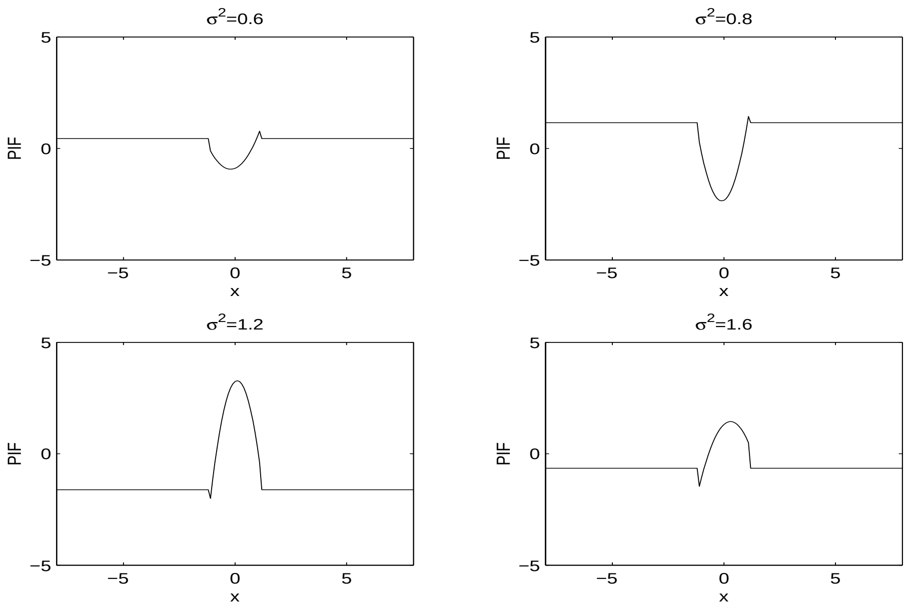

Figure 3: First partial influence function $\mathrm{PIF}_1(x; \mathrm{TPM}_R, H)$*. As in Figure 1, but now using the robust MCD-estimator for estimating the parameters in the discriminant rule* Q*.*

4 Robust Diagnostic Measures and Examples

The heuristic interpretation of (partial) influence functions is that the estimated difference between the population TPM and its estimated value is approximatively given by the average of the values $\mathrm{PIF}(x_i; \mathrm{TPM}, H)$ for $i = 1, \ldots, n$ (cfr. Hampel et al., 1986; Pires and Branco, 2002). Hence the partial influence functions evaluated at the sample points give the contribution of every observation in the training set to the misclassification rate. Large values for the PIF reveal points giving a large positive contribution to the TPM. We restrict ourselves to the detection of influential points for classical discriminant analysis. When a robust discriminant rule Q_R is used, it is less important to pinpoint the highly influential points, since the robust procedure has a bounded influence and is resistant to these observations.

Diagnostic measures are then computed using the first, respectively second, PIF for observations belonging to the first, respectively second, group of training data:

$$\begin{aligned} \mathrm{D}_{i,Cl}(\mu_1, \mu_2, \Sigma_1, \Sigma_2) &= |\mathrm{PIF}_1(x_i, \mathrm{TPM}_{Cl}, H)| \quad \text{for } i = 1, \ldots, n_1 \qquad (4.1) \\ \mathrm{D}_{i,Cl}(\mu_1, \mu_2, \Sigma_1, \Sigma_2) &= |\mathrm{PIF}_2(x_i, \mathrm{TPM}_{Cl}, H)| \quad \text{for } i = n_1 + 1, \ldots, n. \end{aligned}$$

Plotting D_i with respect to the index i, or alternatively with respect to the value of $Q(x_i)$, then results in a diagnostic plot. The sign information in the PIF could be kept by dropping the absolute values in (4.1). To compute the diagnostics D_i, the parameters μ_1, μ_2, Σ_1 and Σ_2 need to be estimated. The prior probability π_1 can be estimated as the frequency of observation from the training sample belonging to the first group, and similarly for π_2.

The idea of using the influence function as a tool for sensitivity analysis has a long tradition in statistics. For applications in multivariate analysis see for example Critchley (1985), and Tanaka (1994). In the construction of the D_i the non-robust sample average and covariance matrix estimators could be used for estimating the population parameters. Though it is well-known that diagnostic measures based on non-robust estimators are subject to the masking effect. Outliers and atypical observations might shift the estimated means and blow up the dispersion matrices, resulting in a non reliable diagnostic measure. It might as well be possible that influential observations will not be detected anymore. To prevent this masking effect, it is proposed to estimate μ_1, μ_2, Σ_1 and Σ_2 using robust estimators, resulting in a robust diagnostic measure. A similar approach to robust diagnostics was taken by (Tanaka and Tarumi, 1996; Pison et al., 2003; and Boente et al., 2003) in different fields of multivariate statistics. In the construction of the robust diagnostic tool, the robust estimators are auxiliary and only serve to estimate the $\mathrm{D}_{i,Cl}(\mu_1, \mu_2, \Sigma_1, \Sigma_2)$ in a reliable way, not suffering from the masking effect. As such, the partial influence function of the non robust classical estimator is estimated in a robust way. The aim is to detect influential points when using Q_{Cl}. When no highly influential points are detected by the robust diagnostic, one could pass to a standard discriminant analysis, the latter one being more efficient at the normal model.

To illustrate the risk of masking when using non-robust diagnostics, consider the *Skull's data*, described in Flury and Riedwyl (1988, page 123-125). This well-known data set contains skull measurements (6 variables) on two species of female voles: Microtus Californicus, and Microtus Ochrogaster. The first group contains 41 observations, and the second 45. In Figure 4 diagnostic plots are made, once using the classical estimators, and once using robust plug-in estimators for $\mathrm{D}_{i,Cl}(\mu_1, \mu_2, \Sigma_1, \Sigma_2)$. The robust diagnostic measures, immediately reveal that there is a huge influential observation: number 73. The non-robust diagnostic measures suffer from the masking effect and cannot detect any influential observations anymore.

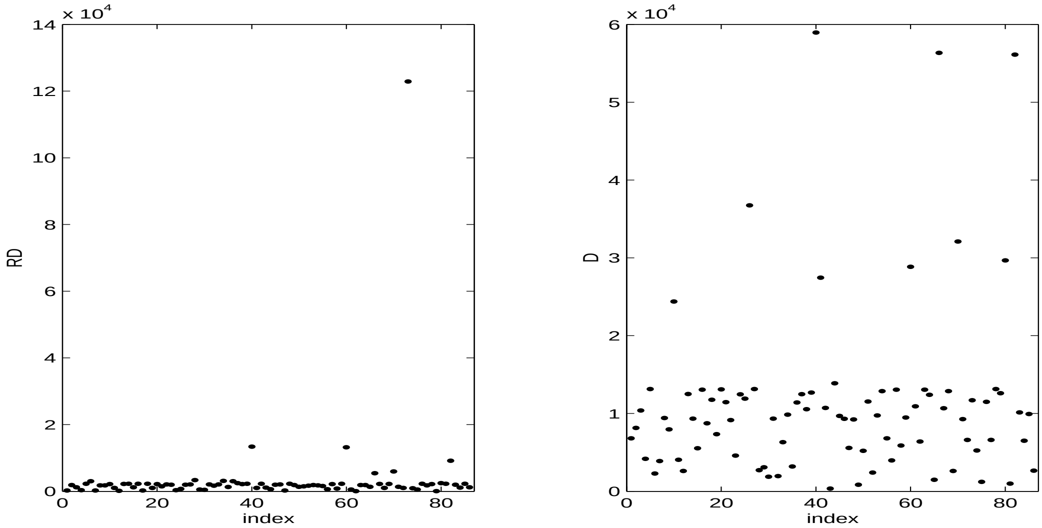

Figure 4: Diagnostic plot for the Skull data using robust plug-in estimators (left figure) or using classical plug-in estimators (right figure) for $D_{i,Cl}(\mu_1, \mu_2, \Sigma_1, \Sigma_2)$.

Several diagnostic measures for classical quadratic discriminant analysis have already been introduced by Fung (1996a). Influence is measured by looking at the effect of deleting an observation from the sample on the estimated probabilities of all other observations. Fung (1996a) proposed different variants, all based on the leave-one-out principle. One of them is the Relative Log-0dds SQuared influence for an observation i,

$$RLOSQ_i = \frac{1}{n}\sum_{j=1}^{n} \left[\log\left\{\frac{\hat{p}_1(x_j)}{1-\hat{p}_1(x_j)}\right\} - \log\left\{\frac{\hat{p}_{1(i)}(x_j)}{1-\hat{p}_{1(i)}(x_j)}\right\}\right]^2,$$

where $\hat{p}_1(x)$ is the estimated probability that an observation x belongs to the first group,

$$\hat{p}_1(x) = \hat{f}_1(x)/[\hat{f}_1(x) + \hat{f}_2(x)],$$

with $\hat{f}_j$ the density of $N_p(\hat{\mu}_j, \hat{\Sigma}_j)$, for $j = 1, 2$. On the other hand, $\hat{p}_{1(i)}(x)$ estimates the same probability, but now using the sample with observation i deleted.

Consider as a second example the *Biting flies* data, described in Johnson and Wichern (2002, page 373). Two species of flies, *Leptoconops cartei* and *Leptoconops torrens*, were thought for many years to be the same, because they are morphologically very similar. For each group a sample of 35 observations was drawn and seven measurements where taken. Figure 6 shows the comparison between the $RLOSQ$-diagnostic and the robust diagnostic based on the partial influence functions for the TPM_{Cl}. The robust diagnostic indicates only 36 as highly influential. The leave-one-out method suggests as well 2, 15 and 23. Further inspection of the data reveals that 2, 15 and 23 are outlying observations. Hence there is a risk that due to the presence of multiple outliers, the leave-one-out procedure becomes unreliable. Whether 2,3, and 15 are highly influential, or only outlying, is difficult to find out using the $RLOSQ$ indices.

5 Conclusions

This paper is about computing the influence of observations in the training sample on the classification error of a discriminant rule. For linear discriminant analysis, answers have been given more than a decade ago, but quadratic discriminant analysis is a harder problem to tackle. Starting from an expression for the total probability of misclassification (Section 2) and using the technology of Partial Influence Functions of Pires and Branco (2002), a computable expression for the influence function was found.

Not surprisingly, this influence function was found to be quadratic and unbounded. Using robust plug-in estimators in the discriminant rule Q, however, yields bounded influence procedures. But it also turned out that whenever the matrix $\Sigma_1\Sigma_2^{-1}$ has eigenvalues close to

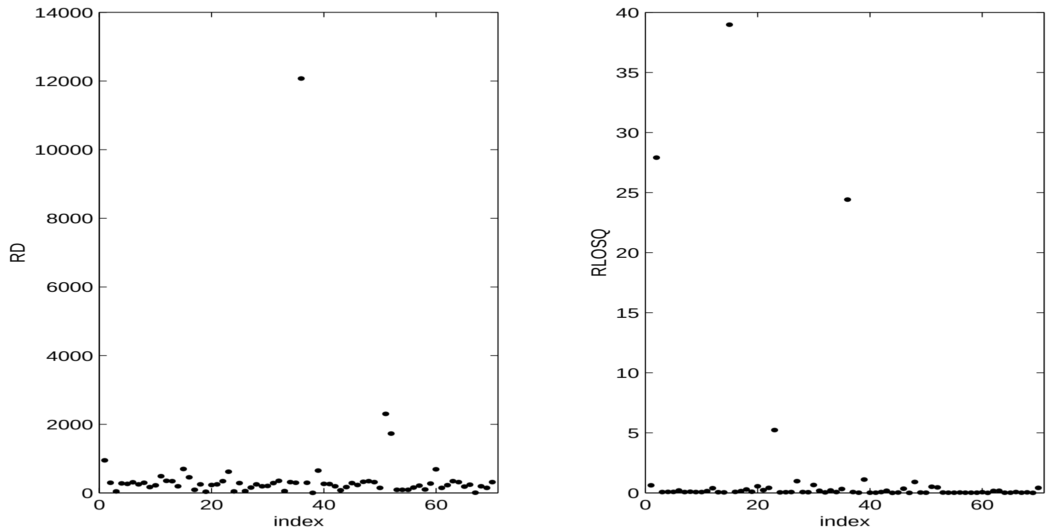

Figure 5: Diagnostic plot for the Biting Flies data using robust diagnostics based on TPM_{Cl} *(left figure) and using the leave-one out measure RLOSQ (right figure).*

each other or close to one, the QDA is unduly sensitive to small data perturbations. Focus was on the influence on the TPM, and not on the influence on the estimates of the parameters of the quadratic discriminant rule. The latter estimates are not of immediate interest in QDA. In some sense, one could think of $\mathrm{PIF}(x; \mathrm{TPM}, H)$ as an appropriate summary of the influences on the estimates of the $p(p+3)$ components of μ_1, μ_2, Σ_1 and Σ_2. Besides of theoretical interest, the PIF can also be used to construct a robust diagnostic tool for the detection of influential points in classical QDA.

Influence diagnostics in discriminant analysis for LDA, QDA, and for the multiple group case were proposed and studied in a sequence of papers by Fung (1995a, 1995b, 1996a, 1996b). In this paper, a theoretical expression of an influence function is used as basis of the diagnostic measure being proposed, allowing to avoid case-wise deletion measures. A completely different approach is taken by Riani and Atkinson (2001), who proposed a forward search algorithm to avoid masking effects in detecting influential points. Their approach is a useful data-analytic tool for a robust sensitivity analysis of a discriminant analysis, and requires user-interactive analysis of the data.

Let us emphasize that we do not aim to develop a new kind of robust discriminant analysis. This paper quantifies the influence of observations on the estimated error rate using plug-in estimates for the parameters of the quadratic discriminant rule. Robust high breakdown linear and quadratic discriminant analysis has been discussed in several papers, such as Hawkins and McLachen (1997), He and Fung (2000), Croux and Dehon (2001), Joossens and Croux (2004) and Hubert and Van Driessen (2004). But most of them focus on computational aspects and simulation comparison. Programs for computing robust linear and quadratic discriminant analysis can be retrieved from *www.econ.kuleuven.ac.be/christophe.croux*.

6 Appendix

Proof of Proposition 1:

It is sufficient to prove (2.7). The quadratic discriminant function (2.1) can be rewritten as written as

$$Q(x; H^0) = (x - \tilde{d}(H^0))^t A(H^0)(x - \tilde{d}(H^0)) - k(H^0), \tag{6.1}$$

with $k = k(H^0)$ defined in (2.9), and $\tilde{d}(H^0) = -A(H^0)^{-1}b(H^0)/2$. Take now $X \sim H_1$, then $W = \Sigma_1^{-1/2}(X - \mu_1) \sim N(0, I_p)$, and definition (2.5) yields

$$\begin{aligned}\Pi_{2|1}(H^0, H) &= P_{H_1}((X - \tilde{d}(H^0)))^t A(H^0)(X - \tilde{d}(H^0)) < k) \\ &= P_{N(0,I_p)}((W - d_{2|1})^t \bar{A}_{2|1}(H^0, H)(W - d_{2|1}) < k),\end{aligned}$$

where $d_{2|1} = d_{2|1}(H^0, H)$ is defined in (2.8). Since $\bar{A}_{2|1}(H^0, H)$ is a symmetric matrix, its eigenvalues λ_j are real and we can write

$$\bar{A}_{2|1}(H^0, H) = \sum_{j=1}^{p} \lambda_j v_j v_j^t,$$

where v_j are the corresponding eigenvectors. Moreover, the eigenvectors of $\bar{A}_{2|1}(H^0, H)$ are orthogonal implying that the variables $W_j = W^t v_j$, for $j = 1, \ldots, p$, are components of a multivariate standard normal distribution.

Proof of Proposition 2:

Equation (3.4) follows from the definition of TPM, and (3.5) results from a standard application of the chain rule. As a first step, the PIF for the estimates of the parameters of the quadratic discriminant rule Q are computed. The matrix derivation rules $\mathrm{PIF}_s(x; \Sigma_s^{-1}, H^0) =$

$-\Sigma_s^{-1}\mathrm{PIF}_s(x;\Sigma_s,H^0)\Sigma_s^{-1}$ and $\mathrm{PIF}_s(x;\log|\Sigma_s|,H^0) = \mathrm{trace}\,(\Sigma_s^{-1}\mathrm{PIF}_s(x;\Sigma_s,H^0))$ for $s=1,2$ are used, cfr. Magnus and Neudecker (1999). Straightforward derivation from definitions (2.2), (2.3), (2.4) yields, then (3.11), (3.12), (3.13).

Since the functional k is a simple combination of the functionals A, b and c, equation (3.8) follows. Lemma 2.1 in Sibson (1979) or Lemma 3 in Croux and Haesbroeck (2000) give influence functions for the eigenvalues and eigenvectors of a symmetric matrix. Applying this result to $\bar{A}_{2|1}(H^0,H) = \Sigma_1^{1/2}A(H^0)\Sigma_1^{1/2}$ results in expressions (3.6) and (3.10). Note that by conditions (i) and (iii), and the fact $\Sigma_1^{1/2}\Sigma_2^{-1}\Sigma_1^{1/2}-I_p$ and $\Sigma_1\Sigma_2^{-1}-I_p$ have the same eigenvalues, division by zero in (3.10) is avoided. From (2.8), equation (3.9) follows and by the definition of d_j^*, equation (3.7) holds for $j=1,\ldots,p$. Of course, similar arguments hold for deriving $\mathrm{PIF}_s(x;\Pi_{1|2},H^0,H)$.

Computation of the partial derivatives of $\Pi_{2|1}(H^0,H)$ *w.r.t.* λ_j, d_j^* *and* k*:*

According to Proposition 1 and with $d_j^* = v_j^t d_{2|1}$, write

$$\Pi_{2|1}(H^0,H) = P\left(\sum_{j=1}^{p}\mathrm{sign}(\lambda_j)X_j^2 < k\right) \quad \text{where} \quad X_j \sim N_p(-d_j^*\sqrt{|\lambda_j|},|\lambda_j|)$$

where the X_j are independent univariate normal variables, each having density

$$f_{X_j}(x_j) = \frac{1}{\sqrt{|\lambda_j|}}\varphi\left(\frac{x_j}{\sqrt{|\lambda_j|}}+d_j^*\right). \tag{6.2}$$

Now (6.2) can be written as the integral

$$\int f_{X_1}(x_1)...f_{X_p}(x_p)I\left(\sum_{j=1}^{n}\mathrm{sign}(\lambda_j)x_j^2<k\right)dx_1\ldots dx_p.$$

By condition (iii) the eigenvalues λ_j of $\bar{A}_{2|1}$ are the same as those of $\Sigma_1\Sigma_2^{-1}-1$ and by condition (i) they are different from zero.

Using the above notations, we get the following three lemmas.

Lemma 1. *The partial derivatives of* $\Pi_{2|1}(H^0,H)$ *with respect to* λ_j *are given by*

$$\frac{1}{2\lambda_j}\left\{-P(\Sigma_i\mathrm{sign}(\lambda_i)X_i^2<k)+E\left[\frac{X_j(X_j+d_j^*\sqrt{|\lambda_j|})}{|\lambda_j|}I(\Sigma_i\mathrm{sign}(\lambda_i)X_i^2<k)\right]\right\},$$

for $j=1,\ldots,p$.

Proof: For each $1 \le j \le p$, it holds that $\frac{\partial}{\partial \lambda_j}\Pi_{2|1}(H^0, H)$ equals

$$\int \frac{\partial}{\partial \lambda_j} f_{X_j}(x_j) \prod_{m=1, m\neq j}^{p} f_{X_i}(x_i) I\left(\sum_{i=1}^{p} \mathrm{sign}(\lambda_i) x_i^2 < k\right) \, dx_1 \dots dx_p$$

$$= \int \mathrm{sign}(\lambda_j) \frac{\partial}{\partial |\lambda_j|} f_{X_j}(x_j) \prod_{m=1, m\neq j}^{p} f_{X_m}(x_m) I\left(\sum_{i=1}^{p} \mathrm{sign}(\lambda_i) x_i^2 < k\right) \, dx_1 \dots dx_p$$

$$\overset{(6.2)}{=} \int \mathrm{sign}(\lambda_j) \left[-\frac{1}{2|\lambda_j|^{3/2}} \varphi\left(\frac{x_j}{\sqrt{|\lambda_j|}} + d_j^*\right) + \left(\frac{x_j}{-2|\lambda_j|^2}\right) \varphi'\left(\frac{x_j}{\sqrt{|\lambda_j|}} + d_j^*\right)\right]$$

$$\prod_{m=1, m\neq j}^{p} f_{X_m}(x_m) I\left(\sum_{i=1}^{p} \mathrm{sign}(\lambda_i) x_i^2 < k\right) \, dx_1 \dots dx_p$$

$$\overset{\varphi'(u) = -u\varphi(u)}{=} \int \mathrm{sign}(\lambda_j) \frac{1}{2|\lambda_j|} \left[-1 + \frac{x_j(x_j + d_j^* \sqrt{|\lambda_j|})}{|\lambda_j|}\right]$$

$$\prod_{m=1}^{p} f_{X_m}(x_m) I\left(\sum_{i=1}^{p} \mathrm{sign}(\lambda_i) x_i^2 < k\right) \, dx_1 \dots dx_p$$

$$= \frac{1}{2\lambda_j} \left\{ -P(\Sigma_i \mathrm{sign}(\lambda_i) X_i^2 < k) + E\left[\frac{X_j(X_j + d_j^* \sqrt{|\lambda_j|})}{|\lambda_j|} I(\Sigma_i \mathrm{sign}(\lambda_j) X_i^2 < k)\right]\right\}.$$

□

Lemma 2. *The partial derivatives of* $\Pi_{2|1}(H^0, H)$ *with respect to* d_j^* *are given by*

$$\frac{-1}{\sqrt{|\lambda_j|}} E[X_j I(\Sigma_i \mathrm{sign}(\lambda_i) X_i^2 < k)] - d_j^* P(\Sigma_i \mathrm{sign}(\lambda_i) X_i^2 < k),$$

for $j = 1, \dots, p$.

Proof: For each $1 \le j \le p$, it holds that $\frac{\partial}{\partial d_j^*}\Pi_{2|1}(H^0, H)$ equals

$$\int \frac{\partial}{\partial d_j^*} f_{X_j}(x_j) \prod_{m=1, m\neq j}^{p} f_{X_m}(x_m) I\left(\sum_{i=1}^{p} \mathrm{sign}(\lambda_i) x_i^2 < k\right) dx_1 \ldots dx_p$$

$$\stackrel{(6.2)}{=} \int \frac{1}{\sqrt{|\lambda_j|}} \varphi'\left(\frac{x_j}{\sqrt{|\lambda_j|}} + d_j^*\right) \prod_{m=1, m\neq j}^{p} f_{X_m}(x_m) I\left(\sum_{i=1}^{p} \mathrm{sign}(\lambda_i) x_i^2 < k\right) dx_1 \ldots dx_p$$

$$\stackrel{\varphi'(u) = -u\varphi(u)}{=} \int \left(-\frac{x_j + d_j^*\sqrt{|\lambda_j|}}{|\lambda_j|}\right) \varphi\left(\frac{x_j}{\sqrt{|\lambda_j|}} + d_j^*\right)$$

$$\prod_{m=1, m\neq j}^{p} f_{X_m}(x_m) I\left(\sum_{i=1}^{p} \mathrm{sign}(\lambda_i) x_i^2 < k\right) dx_1 \ldots dx_p$$

$$= \int \left(-\frac{x_j}{\sqrt{|\lambda_j|}} - d_j^*\right) \prod_{m=1}^{p} f_{X_m}(x_m) I\left(\sum_{i=1}^{p} \mathrm{sign}(\lambda_i) x_i^2 < k\right) dx_1 \ldots dx_p$$

$$= -\frac{1}{\sqrt{|\lambda_j|}} E[X_j I(\Sigma_i \mathrm{sign}(\lambda_i) X_i^2 < k)] - d_j^* P(\Sigma_i \mathrm{sign}(\lambda_i) X_i^2 < k).$$

□

For the partial derivative with respect to k, we will reorder the components of X such that the corresponding eigenvalues satisfy

$$\lambda_{(1)} \ge \ldots \ge \lambda_{(q)} > 0 > \lambda_{(q+1)} \ge \ldots \ge \lambda_{(p)},$$

where q is the number of positive eigenvalues. Furthermore, let

$$S^+ = \sum_{j=1}^{q} X_{(j)}^2 \text{ and } S^- = \sum_{j=q+1}^{p} X_{(j)}^2$$

where empty sums are zero by convention. From (6.2) we have that $\Pi_{2|1}(H^0, H) = P(S^+ - S^- < k)$. Without loss of generality we will suppose that $k > 0$. For $k < 0$ one has

$$\frac{\partial \Pi_{2|1}(H^0, H)}{\partial k} = -\frac{\partial P(S^- - S^+ > |k|)}{\partial |k|} = \frac{\partial P(S^- - S^+ \le |k|)}{\partial |k|}$$

and it suffices to interchange the roles of S^+ and S^- in the lemma below.

Lemma 3. *With this notations above, and for $k > 0$, the partial derivative of Π_{12} with respect to k is given by*

$$\begin{array}{ll} 0 & \text{if} \quad q = 0 \\ E\left[\left\{f_{X_{(1)}}(\sqrt{k+S^-}) + f_{X_{(1)}}(-\sqrt{k+S^-})\right\} / (2\sqrt{k+S^-})\right] & \text{if} \quad q = 1 \\ E\left[\pi^{q-1}(k+S^-)^{\frac{q-2}{2}} f_q(U\sqrt{k+S^-})\delta(\theta(U))\right] & \text{if} \quad q \ge 2 \end{array}$$

where f_q is joint density of $(X_{(1)}, \ldots, X_{(q)})^t$ in polar coordinates, U is uniformly distributed on the periphery of the q dimensional unit sphere S^{q-1}, independently of S^-. Here $\delta(\theta(u)) = \sin^{q-2}\theta_1 \sin^{q-3}\theta_2 \ldots \sin\theta_{q-2}$ for $q \geq 2$, with $\theta(u) = (\theta_1, \ldots, \theta_q)$ the angles determining u.

Proof: The results is clear for $q = 0$ since it was supposed that $k > 0$. Now if $q = 1$ then

$$\begin{aligned}
\frac{\partial \Pi_{2|1}(H^0, H)}{\partial k} &= E\left[\frac{\partial}{\partial k} P(X_{(1)}^2 \leq k + S^-|S^-)\right] \\
&= E\left[\frac{\partial}{\partial k}\int_0^{k+S^-} f_{X_{(1)}^2}(u)du\right] \\
&= E\left[f_{X_{(1)}^2}(k + S^-)\right] \\
&= E\left[\left\{f_{X_{(1)}}(\sqrt{k+S^-}) + f_{X_{(1)}}(-\sqrt{k+S^-})\right\}/(2\sqrt{k+S^-})\right].
\end{aligned}$$

For $q \geq 2$, a transformation $f_q(x_{(1)}, \ldots, x_{(q)}) := f_q(x^q) \to f_q(r, \theta)$ to polar coordinates will be carried out, where $r = \|x^q\|$ and $\theta \equiv (\theta_1, \ \ldots, \ \theta_{q-1})$, with $\theta_1, \ \ldots, \ \theta_{q-2} \in [0, \pi[$, $\theta_{q-1} \in [0, 2\pi[$ contains the corresponding angles. Let Θ be the space where the angles vary in, and let $\theta(u)$ be the set of angles associated with a unit vector. Then $\delta(\theta) = \sin^{q-2}\theta_1 \sin^{q-3}\theta_2 \ldots \sin\theta_{q-2}$ is the absolute value of the determinant of the Jacobian of this transformation. For every positive k one has

$$\begin{aligned}
& \frac{\partial}{\partial k} P(S^+ \leq k) && (6.3) \\
= \ & \frac{\partial}{\partial k}\int f_q(x^q) I(\|x_q\|^2 < k)dx_q \\
= \ & \frac{\partial}{\partial k}\int_0^{\sqrt{k}}\int_\Theta f_q(r, \theta) r^{q-1}\delta(\theta) \ d\theta \ dr \\
\overset{\text{Fubini}}{=} \ & \int_\Theta \frac{\partial}{\partial k}\int_0^{\sqrt{k}} f_q(r, \theta) r^{q-1}\delta(\theta) \ d\theta \ dr \\
\overset{\text{Leibnitz}}{=} \ & \int_\Theta \frac{1}{2\sqrt{k}} k^{\frac{q-1}{2}} f_q(\sqrt{k}, \theta)\delta(\theta)d\theta \\
= \ & \frac{k^{\frac{q-2}{2}}}{2}\int_\Theta f(\sqrt{k}, \theta)\delta(\theta)d\theta, \\
= \ & \frac{k^{\frac{q-2}{2}}}{2} 2\pi^{q-1} \ E_U[f_q(\sqrt{k}, U)\delta(\theta(U))], && (6.4)
\end{aligned}$$

where U is uniformly distributed over the q-dimensional unit sphere S^{q-1}. Then

$$\begin{aligned}
\frac{\partial}{\partial k}\Pi_{2|1}(H^0, H) &= E\left[\frac{\partial}{\partial k} P(S^+ \leq k + S^-|S^-)\right] \\
&= E\left[\pi^{q-1} k_q^{\frac{q-2}{2}} f_q(U\sqrt{k+S^-})\delta(\theta(U))\right].
\end{aligned}$$

$\square$

Finally, it is easy to verify that the partial derivatives of $\Pi_{1|2}(H^0, H)$ with respect to λ_j, d_j^* and k are given by similar expressions as in Lemmas 1, 2 and 3. In Lemmas 1 and 2 the inequalities need to inversed, while the sign of the formula of Lemma 3 needs to be changed.

Acknowledgment: We wish to thank the reviewers for their careful reading of the text and their helpful and constructive comments. This research has been supported by the Research Fund K.U. Leuven and the "Fonds voor Wetenschappelijk Onderzoek" (Contract number G.0385.03).

References

Bensmail, H., and Celuex, G. (1996), Regularized Gaussian discriminant analysis through eigenvalue decomposition, *Journal of the American Statistical Association*, 91, 1743–1748.

Boente, G., Pires, A.M., and Rodrigues, I.M. (2002), Influence functions and outlier detection under the common principal components model: A robust approach, *Biometrika*, 89, 861–875.

Butler, R.W., Davies, P.L., and Jhun, M. (1993), Asymptotics for the Minimum Covariance Determinant estimator, *The Annals of Statistics*, 21, 1385–1400.

Campbell, N.A. (1978), The influence function as an aid in outlier detection in discriminant analysis, *Applied Statistics*, 27, 251–258.

Critchley, F. (1985), Influence in principal components analysis, *Biometrika*, 72, 627–636.

Critchley, F., and Vitiello, C. (1991), The influence of observations on misclassification probability estimates in linear discriminant analysis, *Biometrika*, 78, 677–690.

Croux, C., and Dehon, C. (2001), Robust linear discriminant analysis using S-estimators, *The Canadian Journal of Statistics*, 29, 473–492.

Croux, C., and Haesbroeck, G. (1999), Influence function and efficiency of the Minimum Covariance Determinant scatter matrix estimator, *Journal of Multivariate Analysis*, 71, 161–190.

Croux, C., and Haesbroeck, G. (2000), Principal component analysis based on robust estimates of the covariance and correlation matrix: influence functions and efficiencies, *Biometrika*, 87, 603–618.

Flury, B. and Riedwyl, H. (1988). *Multivariate statistics : a practical approach*, London: Chapman and Hall.

Fung, W.K. (1992), Some diagnostic measures in discriminant analysis, *Statistics and Probability Letters*, 13, 279–285.

Fung, W.K. (1995a), Diagnostics in linear discriminant analysis, *Journal of the American Statistical Association*, 90, 952–956.

Fung, W.K. (1995b), Detecting influential observations for estimated probabilities in multiple discriminant analysis, *Computational Statistics and Data Analysis*, 20, 557–568.

Fung, W.K. (1996a), Diagnosing influential observations in quadratic discriminant analysis, *Biometrics*, 52, 1235–1241.

Fung, W.K. (1996b), The influence of an observation on the misclassification probability in multiple discriminant analysis, *Communications in Statistics - Theory and Methods*, 25, 1917–1930.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986), *Robust Statistics: The Approach Based on Influence Functions,* New York: John Wiley.

Hawkins, D.M., and McLachlan, G.J. (1997), High breakdown linear discriminant analysis, *Journal of the American Statistical Association*, 92, 136–143.

He, X., and Fung, W.K. (2000), High breakdown estimation for multiple populations with applications to discriminant analysis, *Journal of Multivariate Analysis*, 72-2, 151–162.

Houshmand, A.A. (1993), Misclassification probabilities for the quadratic discriminant function, *Communications in Statistics, series B*, 81–98.

Hubert, M., and Van Driessen, K. (2004), Fast and robust discriminant analysis, *Computational Statistics and Data Analysis*, 45-2, 301–320.

Johnson, R. A., and Wichern, D.W. (2002). *Applied Multivariate Statistical Analysis, 4th Edition.* Prentice-Hall, London.

Joossens, K., and Croux, C. (2004), Empirical comparison of the classification performance of robust linear and quadratic discriminant analysis, *Theory and Applications of Recent Robust Methods*, Edited by M. Hubert, G. Pison, A. Struyf and S. Van Aelst, Basel: Birkhäuser, 131–140.

Lachenbruch, P.A. (1979), Note on initial misclassification effects on the quadratic discriminant function, *Technometrics*, 21, 129–132.

Magnus, J.R., and Neudecker, H. (1999), *Matrix Differential Calculus with Applications in Statistics and Econometrics, 2nd Edition,* New York: John Wiley.

Maronna, R., and Yohai, V. (1998), Robust estimation of multivariate location and scatter, *Encyclopedia of Statistical Sciences Update Volume 2*, Edited by S. Kotz, C. Read, and D. Banks, New York: John Wiley, 589–596.

McFarland, H.R., and Richards, D. (2002), Exact misclassification probabilities for plug-in normal quadratic discriminant functions II. The heterogeneous case, *Journal of Multivariate Analysis*, 82, 229–330.

Pires, A.M., and Branco, J.A. (2002), Partial influence functions, *Journal of Multivariate Analysis*, 83-2, 451–468.

Pison, G., Rousseeuw, P.J., Filzmoser, P., and Croux, C. (2003), Robust Factor Analysis, *Journal of Multivariate Analysis*, 84, 145–172.

Randles, R.H., Broffitt, J.D., Ramsberg, J.S., and Hogg, R.V. (1978), Generalized linear and quadratic discriminant functions using robust estimators, *Journal of the American Statistical Association*, 73 , 564–568.

Riani, M., and Atkinson, A.C. (2001), A unified approach to outliers, influence and transformations in discriminant analysis, *Journal of Computational and Graphical Statistics*, 10-3, 513–544.

Rousseeuw, P.J., and Van Driessen, K. (1999), A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, 41, 212–223.

Sibson, R. (1979), Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling, *Journal of the Royal Statistical Society, Series B*, 41, 217–229.

Tanaka, Y. (1994), Recent advance in sensitivity analysis in multivariate statistical methods, *Journal of the Japanese Society of Computational Statistics*, 7, 1–25.

Tanaka, Y., and Tarumi, T. (1996), Sensitivity analysis in multivariate methods: General procedure based on influence functions and its robust version, *Compstat: Proceedings in Computational Statistics*, Ed. A. Prat, Heidelberg: Physica-Verlag, 186–185.